2021-09

DESIGN METHODOLOGIES FOR 21ST CENTURY ENTITY CORRELATION

Watson, Frank O., III

Monterey, CA; Naval Postgraduate School

http://hdl.handle.net/10945/68396

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

DESIGN METHODOLOGIES FOR 21ST CENTURY ENTITY CORRELATION

by

Frank O. Watson III

September 2021

Thesis Advisor:	John Audia
Co-Advisors:	Eric S. Hodson
	Kristin M. Giammarco

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC, 20503.

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** September 2021	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE DESIGN METHODOLOGIES FOR 21ST CENTURY ENTITY CORRELATION	**5. FUNDING NUMBERS**
6. AUTHOR(S) Frank O. Watson III	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

Modern war-fighters use automated tools for provisioning complete and consistent information about entity locations in their operations and threat environment. Next generation software correlation tools will have to process millions of entities per day in order to enhance war-fighter situational awareness. Current software correlation tools fuse entity information using purely mathematical algorithms and methods that do not take into account divergent data source precision, limitations, and origins. As a result of having no framework that accounts for these variables, when multiple, diverse data sources are ingested, the accuracy of a fused entity data deteriorates. This thesis aims to answer the question: How can next generation software tools be improved to provide future war-fighters the most accurate picture of their operations and threat environment possible? New model-based design practices and emerging theories of systems pathology will be used to examine the shortfalls of existing methods. Additionally, by designing enhanced data ingestion models and evaluating them against current methods, this thesis aims to demonstrate how the next generation of software correlation tools can measurably increase the accuracy of entity correlation, improving situational awareness for war-fighters of the future.

14. SUBJECT TERMS entity correlation, track correlation, fusion, data, governance	**15. NUMBER OF PAGES** 71
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

DESIGN METHODOLOGIES FOR 21ST CENTURY ENTITY CORRELATION

Frank O. Watson III
Civilian, Department of the Navy
BS, University of Phoenix, 2009

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN SYSTEMS ENGINEERING MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
September 2021

Approved by: John Audia
Advisor

Eric S. Hodson
Co-Advisor

Kristin M. Giammarco
Co-Advisor

Oleg A. Yakimenko
Chair, Department of Systems Engineering

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Modern war-fighters use automated tools for provisioning complete and consistent information about entity locations in their operations and threat environment. Next generation software correlation tools will have to process millions of entities per day in order to enhance war-fighter situational awareness. Current software correlation tools fuse entity information using purely mathematical algorithms and methods that do not take into account divergent data source precision, limitations, and origins. As a result of having no framework that accounts for these variables, when multiple, diverse data sources are ingested, the accuracy of a fused entity data deteriorates. This thesis aims to answer the question: How can next generation software tools be improved to provide future war-fighters the most accurate picture of their operations and threat environment possible? New model-based design practices and emerging theories of systems pathology will be used to examine the shortfalls of existing methods. Additionally, by designing enhanced data ingestion models and evaluating them against current methods, this thesis aims to demonstrate how the next generation of software correlation tools can measurably increase the accuracy of entity correlation, improving situational awareness for war-fighters of the future.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AIS	automated identification system
DOP	dilution of position
MDA	maritime domain awareness
MHz	megahertz
MP	Monterey Phoenix
NIST	National Institute of Standards and Technology
OTH-Gold	over the horizon-gold revision
SysML	systems modeling language
TADIL	tactical digital information link
TDMA	time division multiple access
UML	unified modeling language
USCG	United States Coast Guard
USN	Unites States Navy
VHF	very high frequency

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The next generation of battlespace correlation and fusion tools will have to process millions of battlespace objects each day from divergent data sources to enhance warfighter situational awareness. As the diversity and quantity of data sources increases, human analysts must expend significant effort and time to filter and discriminate unwanted data from correlator displays to create more accurate depictions of entities in the battlespace. Any means of relieving unnecessary processing burden, reducing software processing latency, and consuming less memory could be expected to improve both correlator and human analyst performance.

The extreme volume of daily Automated Identification System (AIS) track reports that are ingested today creates an immense correlator workload comparable to the extreme number of tasks associated with big data analysis. This workload includes processing corrupted position reports and evaluating all relevant position report data fields from good data sources, increasing the amount of valuable system computational resources for every received report. In light of this problem, this research asked the question, "How can behavior models of divergent data sources be used to design improvements in accuracy of battlespace object correlation and fusion?"

A design and analysis research method was used to identify and explore algorithm-based multi-source fusion models and to compare ungoverned and governed approaches to fusion. The research was conducted in three phases. The first phase modeled and evaluated current methodologies in use today. The second phase defined a new data governance-based approach to the problem, and the third phase integrated a hybrid methodology that implemented data governance, determining the extent by which results improved. The implementation of a formal modeling approach that demonstrates expected and unintended behaviors of complex system data flows provided unique insights not available via informal methods. This modeled information provided a framework for implementing data governance, which has the potential to significantly relieve the processing burden from correlation software while also reducing the burden of human analysts correcting correlated visualizations of ship positions.

Modeling event flows in Monterey Phoenix (MP) proved to have a value by itself, as it articulated both the current and future processes of data correlation and fusion in the context of the entire system (correlator) and the environment (data sources and human analysts). Modeling existing AIS report data flow using MP demonstrated divergent behaviors of ungoverned data sources, identifying high levels of AIS data source corruption. Creation of an AIS report data governance framework addressed this issue by applying discrimination factors evaluating the quality of unique data sources prior to correlation. Formal behavioral modeling of data sources that contain statistical variance provided the ability to create and analyze the data governance framework via analysis of MP event traces. Tactical information flow modeling with the data governance framework applied discrimination factors, or "rubrics" for data sources prior to processing. The implementation of the data governance framework that discriminates and filters corrupted data from position report sources prior to sending those reports to an AIS correlator decreased the probability of tactical correlators processing corrupted AIS position reports.

The analyses in this research established that sending large quantities of corrupt position reports into AIS correlator software represents "futile" processing cycles in tactical correlation software. The formal models developed in MP helped articulate the concept of a governance framework and quantitatively demonstrated the impact to the accuracy of the fused picture of self-reporting AIS vessels in the battlespace. The current, ungoverned data architecture in use today provides only a 2.94% probability (based on the data set used) that all AIS reports sent to be correlated were completely usable, or uncorrupted. The implementation of a data governance framework on the same data set significantly increased the probability of completely usable, or uncorrupted AIS reports being sent for processing, increasing that probability from only 2.94%, to 98.01%.

This research validated the essential value of data governance and demonstrated the application of a data governance framework as a simple, viable method of discriminating data source quality. Tactical correlation software runs more efficiently, decreasing resource consumption while providing a more accurate depiction of the battlespace, increasing the consistency and accuracy of data sources. Future correlation software and human analysts will be more efficient when receiving uncorrupted data. Human analysts will be able to invest more time and mental focus into examining entities and interactions rather than expending energy and focus using historical context and other methods to purge incorrect position reports from their three-dimensional understanding of the battlespace.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGEMENTS

I would like to acknowledge the amazing work of my peers in this field. We are nothing without the knowledge of our professors and mentors, who help us navigate the chaotic and exponential increase of data sources in an already saturated combat information environment. I would specifically like to thank my Naval Postgraduate School (NPS) advisors. Kristin Giammarco, John Audia, and Eric Hodson provided critical insight and guidance throughout this research. Without their collective expertise and wisdom, this thesis would not have been possible.

I would especially like to thank my family. My wife, Anita, and my daughter, Kiana, have been extremely supportive of me throughout this entire process and have made countless sacrifices to help me get to this point.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

This thesis investigates the use of model-based design practices for classifying the fidelity and accuracy of divergent data sources used in automated object tracking. Existing methods of automated object tracking, also called battlespace object correlation and fusion, fall victim to data source deviation over time. This research investigates how these current methods propagate aggregated data inaccuracies when correlating multiple battlespace data sources simultaneously, and it proposes a solution in the form of an applied data governance framework.

A. UNIQUE CHALLENGE

The proliferation of information in the 21st century represents a unique challenge to modern warfighters. Every year, increasing numbers of smart platforms, airframes, vehicles, and sensors are fielded. Many of these platforms and sensors report their location in geospatial space and time, or report objects detected in geographic location and time. In order to provide situational awareness of the battlespace, all of these reporting and detected objects must be correlated together to create a fused picture of the battlespace. Objects either report themselves, or are reported by sensors that detect them, and then standardized messages are shared via broadcasts. National subscription services re-broadcast aggregated track message data to tactical users via discrete networks that send these messages to tactical correlators or simple informational displays to visualize the information to provide situational awareness or maritime domain awareness to warfighters. Current battlespace object correlation and fusion tools track objects using purely mathematical algorithms and methods that do not reconcile divergent data source precision, limitations, and data origins.

B. RESEARCH QUESTION

How can behavior models of divergent data sources be used to design improvements in accuracy of battlespace object correlation and fusion?

C. MOTIVATION

Today's warfighters use automated tools for provisioning complete and consistent information about hundreds of thousands of battlespace objects or tracks in their operations and threat environment.

The concept of information corruption and overload is not limited to news or social media today. It applies to characterization of the battlespace as well. For example, if you had a group of witnesses all reporting criminal behavior of a certain person, how would you determine who committed the crime, where they committed the crime, and which witness was giving you the most accurate information? The answer to this question is nuanced. As lead investigator, one would interview the witnesses to determine their observation ability, accuracy, and location in reference to the crime. Determination is made on how well a witness perceived the crime, and if their observations were biased in any way. An investigator would do this to determine how their biases affect their perception. In other words, an investigator would review the sources of the data, and apply disciplined data governance to those sources to filter out unnecessary witnesses, biased witnesses, and false witnesses from a determination of the criminal's identity, and where the crime was committed. This concept also applies to self-reports. If a person reported themselves to the investigator for a crime, the investigator needs to determine if the person is under the undue influence of someone else, drugs, or disabled in some capacity to understand their environment. Globally, this is the exact decision-making process of data governance that must be applied to correlation of object tracks in the battlespace. Unfortunately, the burden of this decision-making process is largely the workflow of human operational intelligence analysts who man 24/7 watch floors, spending numerous hours of their watches, merging, deleting, and fusing tracks together to clean up the common operational track presentation of the battlespace.

Without a governing framework to account for these variables when multiple, diverse data sources are simultaneously ingested, the accuracy of fused battlespace object data deteriorates. Human intelligence analysts are forced to manually review and correct the track picture that is created. The next generation of battlespace correlation and fusion tools will have to process millions of battlespace objects per day from many divergent data sources to enhance warfighter situational awareness.

D. RESEARCH METHOD

This thesis utilizes a design and analysis research method to identify and explore algorithm-based multi-source fusion models. The first step of this research is to conduct a literature review to examine academic articles, journals, dissertations, conference proceedings and other resources relevant to maritime domain awareness, correlation and fusion, and data governance issues. This review provides practical perspectives for this thesis by identifying past research in this topic.

The remaining research is conducted in three phases. The first phase models and evaluates current methodologies in use today. The second phase defines a new data governance approach to the problem, and the third phase integrates a hybrid methodology that includes data governance that informs a determination about the extent results are improved.

The first phase of research involves constructing a model of a current common multisource fusion methodology that will establish a baseline showing the current error rates in standard approaches. Existing mathematical algorithms are used to identify objects using limited parameters such as geospatial location, and time of report. Humans currently use historical norms and pattern analysis to provide perform error corrections that improve track accuracy, or what we refer to as data governance. We use the Naval Postgraduate School Monterey Phoenix tool to develop this model. Using critical source attributes, we create a parameterized scoring rubric that applies source accuracy ratings in Monterey Phoenix. This notional scoring system evaluates data sources for corruption, which adversely affects data source accuracy. The scoring system is based on statistical analysis of the report sources and applies standardized statistical benchmark ratings to reporting sources. Statistics from three real systems labeled A, B, and C are used to characterize the behavior models. Evaluations of these three sources collate information source accuracy ratings into a data source accuracy rubric. We render event traces, flow diagrams and reports from MP, and produce tables to display model statistics pertaining to specific event traces and across all of the event traces. Analysis of the model identifies the shortcomings of fusing object tracks that are based solely on mathematical algorithms.

The second phase of research describes data governance in relation to fusing object tracks. Data governance is defined and evaluated against current fusion methodology to determine which aspects of data governance might be established as a framework.

The third phase of research is focused on creating a hybrid approach that integrates automated data governance. It is an MP model that automates elements of track position accuracy analysis that would otherwise be done by human analysts. Data sources are modified with data governance rubric values and root event behaviors to provide data governance automation in this model. As in the first stage, data source evaluations are conducted. Source quality attributes are integrated into an overall position accuracy rubric. We again render event traces, flow diagrams and reports from MP, and produce tables to display model statistics. A comparative analysis of the event traces is conducted against the models with and without automated data governance. In particular, differences in the data source corruption levels are measured to identify any shortcomings in either model with respect to accuracy based on existing data sets. Analysis is conducted on the results to verify the extent to which (if any) the hybrid model improves data source accuracy vs the current method of common multisource fusion.

E. BENEFIT OF THIS STUDY

This research and the algorithm developed in the form of an MP model for behaviors, informs design improvements to the next generation of battlespace object correlation and fusion tools and methodologies. The ultimate goal is to measurably increase the accuracy of battlespace informational displays and situational awareness by providing a guide to improving current methods, which in turn will improve combat situational awareness for future warfighters.

F. THESIS ORGANIZATION

This chapter focused on the initial problem statement, motivation, and scope of this research. The next chapter establishes research requirements and presents the literature review. Chapter III develops Monterey Phoenix models used to present and study reporting data flow behaviors and conducts an in-depth analysis of model behaviors. Chapter IV provides the conclusions of this research, presents the limitations encountered, and provides recommendations from this thesis for future research. Annotated source code of models created for this research are provided in the Appendix.

THIS PAGE INTENTIONALLY LEFT BLANK

II. LITERATURE REVIEW

There have been many attempts to develop the perfect sensor algorithm that evaluates all possible types of sensor data and entity data from the surface, sub-surface, air, and space domains, but every attempt fails to grasp the scope of the problem. There are popular algorithms in use in the industry today, such as the Kalman filter algorithm, (Wilner 1976) Bayesian fusion algorithms that seek to solve the problem of tracking multiple maneuvering targets from missing and false measurements, (Blom 2006), the maritime traffic knowledge discovery and representation system algorithm, (Fernandez 2018) and many others too numerous to detail in this research thesis. Every day, commercial entities and the military industrial complex are building more divergent and unique sensors and self-reporting systems than ever before. As these systems deploy, complexity of parameters and track attributes as well as data corruption increases. One might intuit that a single complex mathematical algorithm or methodology cannot account for these un-planned emergent behaviors created by old sensors with large inaccuracies as well as new types of reporting capabilities in the multi-domain battlespace. Recent examples of emergent behavior include AIS satellites that are unable to process all the reports from merchant vessels due to saturated communications and processing backlogs. Commercial and military national distribution systems can inadvertently transmit corrupt and time late reports due to mischaracterization of data types and data backlogs.

There are many studies referencing Naval Maritime Domain Awareness (MDA). However, since the focus of this research is on the application of data governance frameworks to multisource fusion, these studies are not reviewed in detail and are only referred to as appropriate. A list of definitions that are specific to this area of research is below.

A. DEFINITIONS

This list of definitions is provided to assist readers in understanding terms related to this research. The specific focus area of this research is Maritime Domain Awareness (MDA).

1. Correlation

For this research, "correlation" refers to the process of determining if mutual relationships or connections exist between two or more battlespace objects. Correlation is "the process by which sensor measurements and other information are combined to keep the [object's position] up-to-date in real time" (Dietrich 2001, 8).

2. Fusion

The term "fusion" is the concept of merging two objects with similar or identical traits into one battlespace object. "Fusion is the process of taking a new a new input (called a contact), comparing it to a database of previous inputs (called tracks), and deciding whether the new input is updated/revised information about an existing track or is a new, previously unreported input that should be added as a new record in the database" (Dietrich 2001, 31).

3. Multi-source Fusion

The term "Multi-source Fusion" in this paper refers to the process of combining sensor and report data from divergent sources to create a refined estimate of an entity's location in time and space. "Data fusion technology structuralizes and integrates heterogeneous data from different sources which greatly improves the comprehensiveness, availability and extensibility of data" (Weiming Liu 2019, 1).

4. Battlespace

The term "battlespace" includes the complete three-dimensional areas of air, sea, and land where fused objects are meant to be displayed to enhance the situational awareness of warfighters (Collins Dictionary 2021).

5. Track

For the purposes of this research, the term "track" represents a detected object or self-reporting entity in the battlespace. Tracks can be aggregate and represent the detected or reported movement of an entity or object in the battlespace over time.

6. Fused Track

For the purposes of this research, the term "fused track" represents an object or self-reporting entity in the battlespace that has been refined by combining and comparing multiple reports using fusion correlator software, with the goal of improving or refining the accuracy of the track's position in time and space using data from multiple sources.

7. Attributes

For the purposes of this research, the term "attributes" refers to standardized fields or parameters of information that are sent in messages from either sensors or self-reporting objects in the battlespace via standardized distribution methods, or the message format of automated tracking and reporting systems that disseminate data across the battlespace.

B. PLATFORM GPS -DILUTION OF POSITION

The most common measurement errors in GPS are cause by geometric Dilution Of Position (DOP). The positional accuracy of automated GPS position reporting is based on the number of satellites a GPS receiver can receive signals from, the relative geometry of those satellites in relation to each other, and the sensitivity of the receiver. Additionally, there can be tropospheric and ionospheres' effects on the GPS signals sent from satellites. Below is a basic GPS pseudo-range formula.

$$P = p + c(dT - dt) + d_{ion} + d_{trop} + e$$

[P] represents the pseudo-range measurement; [p] is the distance between the satellite antenna and the surface antenna receiving the GPS signal; [dT]is receiver clock offset; [dt] is satellite clock offset; [d_{ion}] represents ionosphere propagation delay; [d_{trop}] represents troposphere propagation delay (Langley 1999). This formula shows how critical clocking and distance measures are for accurate position reporting. The average position area of uncertainty for commercial GPS receivers used for AIS reporting is 25 meters. An example DOP for a stationary AIS transponder is shown in Figure 1. The area of uncertainty for a stationary vessel sending AIS reports is approximately twenty meters as demonstrated by the historical divergence of reporting position from an AIS transponder on a stationary vessel sending 94 AIS reports over a 24-hour period.

Figure 1. Dilution of Position

C. AUTOMATED IDENTIFICATION SYSTEM REPORTING VOLUME ISSUES

The Maritime Automated Identification System (AIS) evolved from commercial GPS implementation with global implementation circa 2002, as a method for ships to broadcast their GPS positions and identity via VHF radio frequency to other ships within radio frequency range in order reduce and avoid collisions during inclement weather conditions and after dark. This technology quickly gained popularity as it reduced accidents in congested navigable waterways, straights, and shipping lanes. Currently, in the United States, any class-A commercial ship is required to transmit AIS on VHF frequencies at 161.975 MHz or 162.025 MHz, also known as marine channels 87B, and 88B (United States Coast Guard 2021). In 2009, commercial satellites were launched with the ability to receive the VHF AIS signals from space. These satellites created the capability to globally capture AIS position reports from ships across the globe. Therefore, the ability to process and distribute AIS position reports from across the globe was created. The popularity of AIS increased exponentially between 2010 and 2020, when GPS technology became mainstream and inexpensive. Affordability allowed most modern maritime vessels to deploy GPS based AIS transponders. The explosion of transponders world-wide has caused bottlenecks in satellite receiving capability due to the use of Time Division Multiple Access (TDMA) technology used in the receivers. The TDMA only allows for a satellite or ship-based radio to receive 4500 reports per minute (United States Coast Guard 2021). The busiest shipping lanes in the world overload the TDMA processing capability of AIS receivers, causing missing and garbled ship's position reports.

Modern correlation and fusion software is capable of filtering corrupt reports from in depth analysis but must still receive and ingest every raw AIS report that is received. As displayed in the below graphic, over 500 million AIS reports are generated every day world-wide (Perobelli 2016).

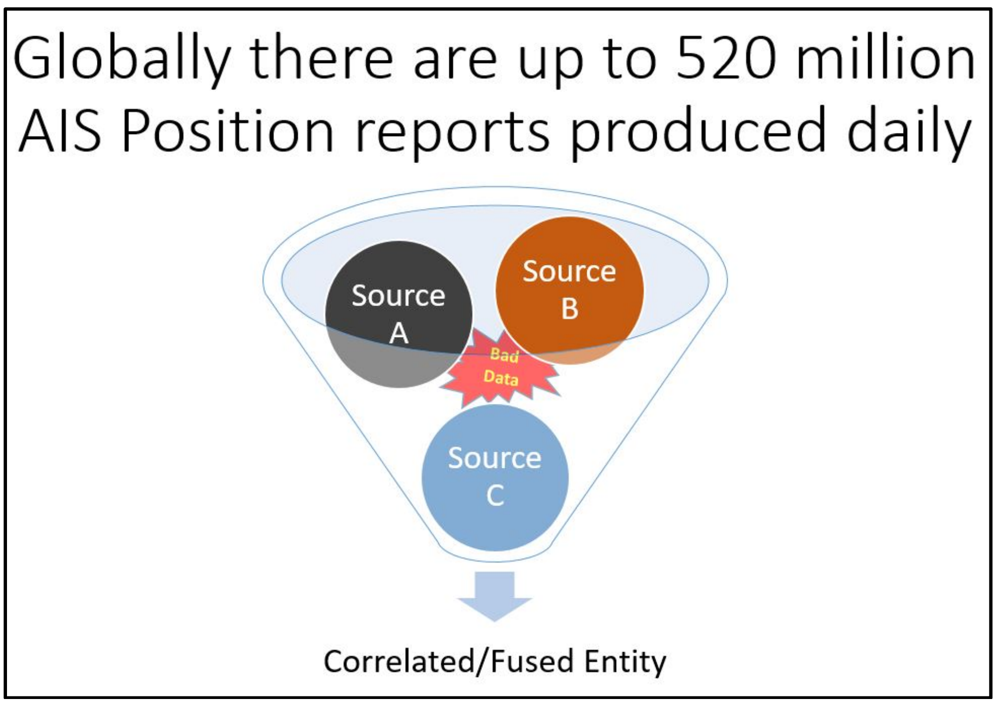

Figure 2. Global AIS Reporting

Due to the volume of reports that are processed, bad data or corrupt reports with missing critical values increase the analysis workload of AIS report correlation causing symptoms including, long processing cues, message processing errors, impeding software correlators processing capabilities and the timeliness and accuracy of AIS report position analysis. Per classical fusion theory, the highest importance is placed on Correlating the identity of an entity, evaluating the temporal accuracy of the AIS report on the entity, and determining geospatial accuracy of the entity's location in the report. The below Venn diagram shows the relationship between the three major focuses of AIS report evaluation.

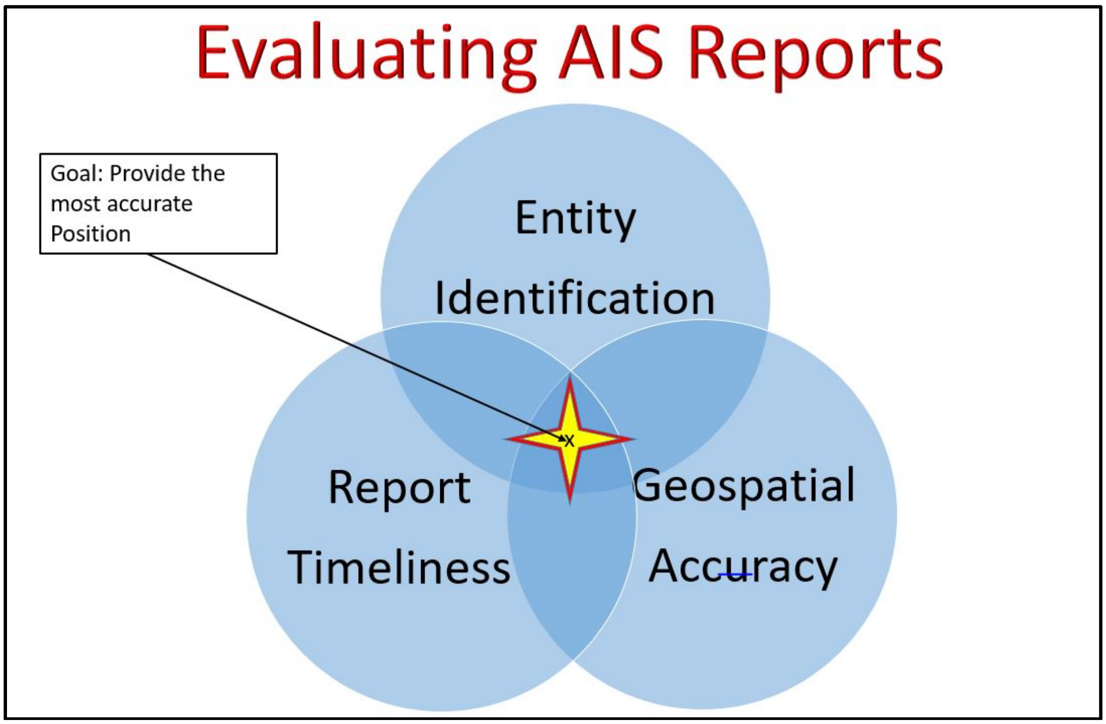

Figure 3. Evaluating AIS Reports

D. DATA GOVERNANCE

Data governance in terms of this research, is the application of higher level decision making and strategy to raw data. How does this apply to creating a three-dimensional picture of the battlespace? The raw collection of tracks and messages, and the process of receiving those tracks from divergent sources into a single battlespace correlator requires data governance. "Attempting to fuse data from a set of heterogeneous systems is a challenge as each system [or source] may define its data objects differently" (Rothenhaus 2008, 17). Inaccurate tracks, old tracks, and multiple reports about the same objects in space from different sources must be intelligently assessed and fused to create a clear and accurate view of the battlespace. The next section focuses on how data governance should be applied to all track messages proliferated throughout commercial industry and the military fleet today. The National Institute of Standards and Technology (NIST), defines data governance as “the over-arching discipline, administration, and formalization of data management” (National Institute for Standards 2019). An example of just one area of this problem is the global shipping community. The national requirement of all commercial vessels 65 ft. in length or larger to have AIS transponders while having different brands of transponders and qualities of Global Positioning Signal (GPS) locational navigation aids is improving reporting rates, but the accuracy of older technologies is problematic. Aging vessels have maintenance issues causing failure to properly report their ship’s position, while newer vessels automatically report their position so often they can overload AIS tracking systems ability to simultaneously correlate large numbers of self-reporting vessels in a crowded commercial port. The confines of restricted navigable waterways with high ship throughput compounds this issue further. These concerns among others have created a constantly growing data management and accuracy problem. Correlating position reports from these vast oceans of data from divergent AIS sources in the most accurate manner is an impossible challenge without a disciplined application of data governance.

For research purposes, this thesis will focus on multi-source AIS data governance, and how apply it, and model its supporting correlation and fusion of AIS track reports. When applying data governance, one must define standards. Those standards must be reasonable and intuitive (National Institute for Standards 2019). These standards create extensible, explainable evaluations of data sources. Once applied they create a data governance framework for information from divergent sources that warfighters have no control over. The graphic below displays the standard AIS message format and fields for class-A AIS reports (United States Coast Guard 2021). There are 15 standard message fields that are correlated with the intent of creating an “improved” fused entity.

Figure 4. Multiple Report Sources

E. APPLYING BEHAVIOR MODELING TO THE DATA GOVERNANCE PROBLEM

Analyzing the problem from a disciplined, system engineering perspective, creates opportunities to look at what has historically been a physics problem in a new light. Modeling and simulation allows system engineers the unique ability to visualize and identify unexpected problems in systems design and data flow. Solutions to identified problems can be evaluated within the virtual models as analogs of production systems to save significant time and investment by evaluating systems design prior to building prototypes.

Three modeling methods were considered for this thesis. Unified Modeling Language, (UML) is intended for creating visual representations of systems and focusing on levels or layers of abstraction. This language helps visualize and document the structure and design of systems. Unified Modeling Language is an excellent tool for mapping requirements but is not executable to evaluate design architecture. There are no methods provided in UML architecture for evaluation emergent behaviors of data sources. Systems Modeling Language, (SysML) is a general-purpose modeling language for systems engineering applications, was considered as well. "SysML is a dialect, or a profile of UML 2.0. UML and SysML provide a coherent language for system diagrams; however, these definitions are not intended to assess the source content of the architectures that they represent" (Quartuccio and Giammarco 2018, 11). While SysML offers the ability to trace functional flow of systems, creating flow arrows, it focuses on dependency sequences built on success or failure of predecessor functions or operations. The Naval Postgraduate School's Monterey Phoenix tool was similarly considered. "Monterey Phoenix is a high-level, executable language for expressing complex event behaviors using a simple event grammar" (Auguston and Giammarco 2019 ,3). Monterey Phoenix helps evaluate planned behaviors of data flow architecture and excels at finding emergent behaviors and system design shortfalls. Modeling of this type differentiates itself from other approaches by demonstrating divergent behaviors of complex systems and data flows. All possible event interaction scenarios and behaviors are created in traces. These event traces provide critical insight into evaluating information flow and models, providing exhaustive trace generation up to a user-defined scope limit (Auguston and Giammarco 2019, 6). These capabilities made MP the best choice for conducting this data governance analysis. For this thesis, MP is employed to describe how errors in the data flows or sources create unwanted behaviors when multiple sources of differing fidelity are correlated into a single object or entity. Corrupted data sources, as well as source time, deviation, geospatial location, and data source abnormalities are the focus of this research.

THIS PAGE INTENTIONALLY LEFT BLANK

III. MODELING MANUAL VERSUS AUTOMATED DATA GOVERNANCE

The purpose of modeling is to provide abstract representations of systems. These abstract representations provide unique insight, supporting specific areas of interest in the field of systems engineering. Models can be in almost any format as long as they provide engineers and designers better understanding of system, while providing the potential to identify design problems. Models provide the best value early in system design processes, before prototypes are built as they can immediately capture underlying design flaws that can be fixed before prototypes are built. Systems engineers use models to link concepts to actual designs, presenting insights to the larger group of product or system stakeholders. The Monterey Phoenix tool will be used to build models and evaluate the modeled event deviation behaviors and the sources of the deviations. Once the deviations are identified, a data governance model will be created that accounts for and evaluates track source accuracy. Once source accuracy deviation is modeled, the attribution of source derived data deviations will be incorporated as an applied data governance framework prior to multi-report correlation.

A. AIS REPORT DATA SOURCES

Data integrity is critical to creating an accurate assessment of entities in the battlespace. Corrupted or missing data fields in reports render these reports useless for analysis when critical fields or data values are missing. While the AIS communications standard contains 15 typical message fields, the most relevant values for report correlation are below.

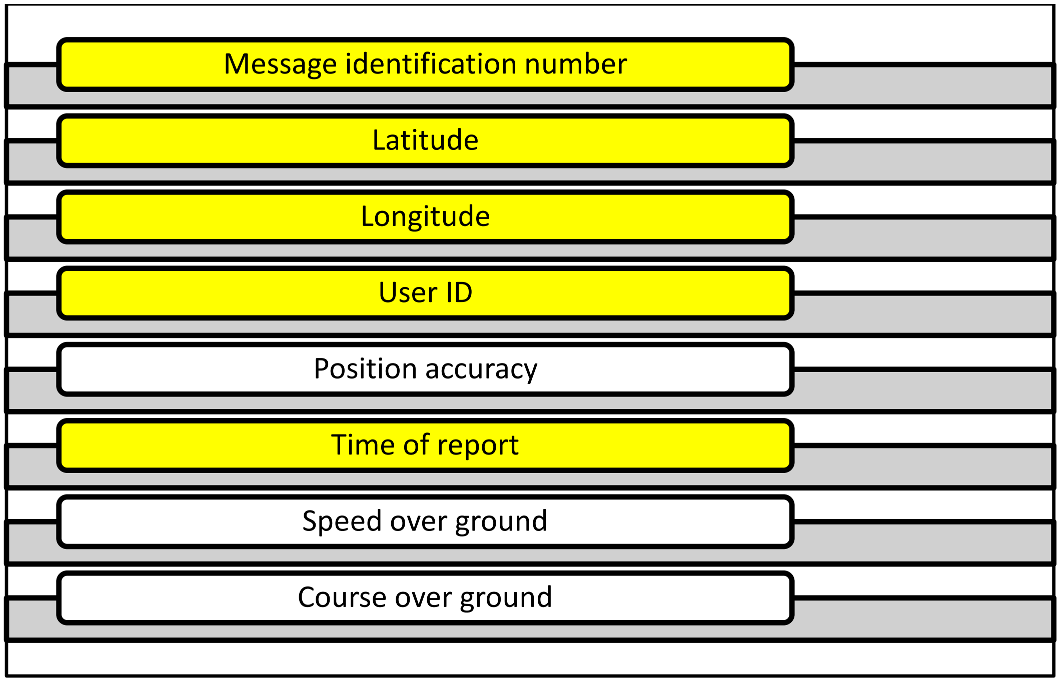

Figure 5. Relevant Values for Report Correlation (with Critical Fields Highlighted)

Five critical fields for correlation analysis are highlighted in yellow. If any of these fields are missing from the report, a standard situational awareness correlator will process the position report, evaluate it, classify it as corrupted, and scrub the report from the local track database, or in the worst case, attempt to merge it with another report that is not corrupted. Every corrupted report received creates processing tax on the situational awareness correlator, and if the correlator merges it with good data, massive positional reporting inaccuracies are created. Without any of the below fields, the report is unable to be processed by the correlator's logic.

Three separate AIS report sources were provided by the United States Coast Guard (USCG) for analysis. These reporting sources and their effects on report correlation were modeled in MP. The data sources have been de-identified for the purposes of this research. The data sources are labeled as data source A, source B, and source C, for examination.

B. IDENTIFYING METRICS AND MEASURES FOR ACCURACY

Metrics and measures establish baseline accuracy statistics of the different data sources. Each of the data sources was statistically evaluated in two different ways. The first method evaluated the proportion of reports from the source that were corrupted against the proportion of the data source that was not corrupt.

Below is the standard proportion formula used in this analysis. The symbol "ρ" represents proportion. Epsilon, "ϵ", represents reports with errors, or corrupted reports. Dividing "ϵ" by the total received reports for a specific data source gives the value of "ρ" . This "ρ" value represents the proportional percentage of reports.

$$\rho = \frac{\epsilon}{Total\ Reports_{Source}}$$

For statistical evaluation of all sources combined, the proportional formula used is below. Epsilon reports with errors of sources A, B, and C are summed up and divided by the total number of reports from all three sources providing "ρ", the proportional percentage of all reports with errors or corruption.

$$\rho = \frac{\epsilon_A + \epsilon_B + \epsilon_C}{Total\ Reports_A + Total\ Reports_B + Total\ Reports_C}$$

The table below displays the data corruption statistics of the three AIS reporting sources provided by the USCG for analysis. For the purposes of this research, the data sets were limited, representing one year of AIS data, collected between January and December of 2019. Analysis was limited to United States flagged research vessels transmitting AIS signals.

Table 1. Data Corruption of Sources

Report Source	% of Valid Reports (100 - ρ)	% of Invalid reports ρ
A	100	0
B	100	0
C	2.66	97.33
All sources Combined	63.43	36.57

Sources A and B showed no measurable corruption, with all reports containing valid user identification (vessel name), message identification number, time of report, latitude, and longitude. Source C results, on the other hand, had significant issues. 97.33% of all reports from source C were corrupted with missing or corrupted values in valid user identification (vessel name), message identification number, time of report, latitude, and longitude. The corruption of data source C impacts the combined report quality from all sources containing good AIS position reports, significantly dragging down the "all sources combined" statistics for valid reports. This results in the insertion of high quantities of corrupt AIS position reports into correlation workflows.

Analysis of USCG AIS data for United States flagged research vessels demonstrates that over 36% of all position reports provided from the aggregated data set including data sources A, B, and C, are corrupted as summarized in Figure 6.

		Data Source A			
Total # Corrupted reports	0	Total # Uncorrupted reports	283	total	283
% Corrupted Reports	0 %	% Uncorrupted reports	100 %		
		Data Source B			
Total # Corrupted reports	0	Total # Uncorrupted reports	1523	total	1523
% Corrupted Reports	0 %		100 %		
		Data Source C			
Total # Corrupted reports	1058	Total # Uncorrupted reports	29	total	1087
% Corrupted Reports	97.3321 %		2.66789 %		
% OF CORRUPTED DATA FROM ALL SOURCES	36.571 %	TOTAL REPORTS PROCESSED BY CORRELATOR			2893

Figure 6. Evaluation of Data Source Corruption

C. MODELING UNGOVERNED DATA FLOW

An initial Monterey Phoenix model was created to show the current data flows used to send AIS reports to tactical correlators today. This model formally lays out all logical steps of data flow, from the reporting source, all the way to the human analyst. The inclusions of environment activities are important as they bring visibility to the sequence of events that typically happen for correlation and fusion. This section first discusses the typical event flows including a sequence of events that represents the baseline case. Next, it discusses the characterization of events with probability attributes, and how Monterey Phoenix mathematically evaluates the likelihood of all possible outcomes.

1. Generation of Typical Event Flows

The event trace shown in Figure 7 depicts how data sources are processed by track correlators today. Starting at the AIS Sensor, reports are sent to AIS aggregation services, and input into global distribution mechanisms. Reports from various sources are then sent to tactical situational awareness correlators to be processed. In the MP model, data flow is modeled starting at the sensor, flowing to the aggregation service, then routing to the distribution mechanism. All three USCG data sources are represented and transmitted to the correlator. AIS reports are processed and evaluated before being sent to the display for analyst viewing. The Monterey Phoenix model was run using scope "1", which generates sufficient modeling traces and the likelihood of all outcomes for analysis. Eight possible results were traced using Monterey Phoenix. Each trace displays the percentages of good versus corrupted data sources on the top left of the figure. Figure 7 represents the optimal trace with the outcome of receiving "un-corrupted" reports from all data sources.

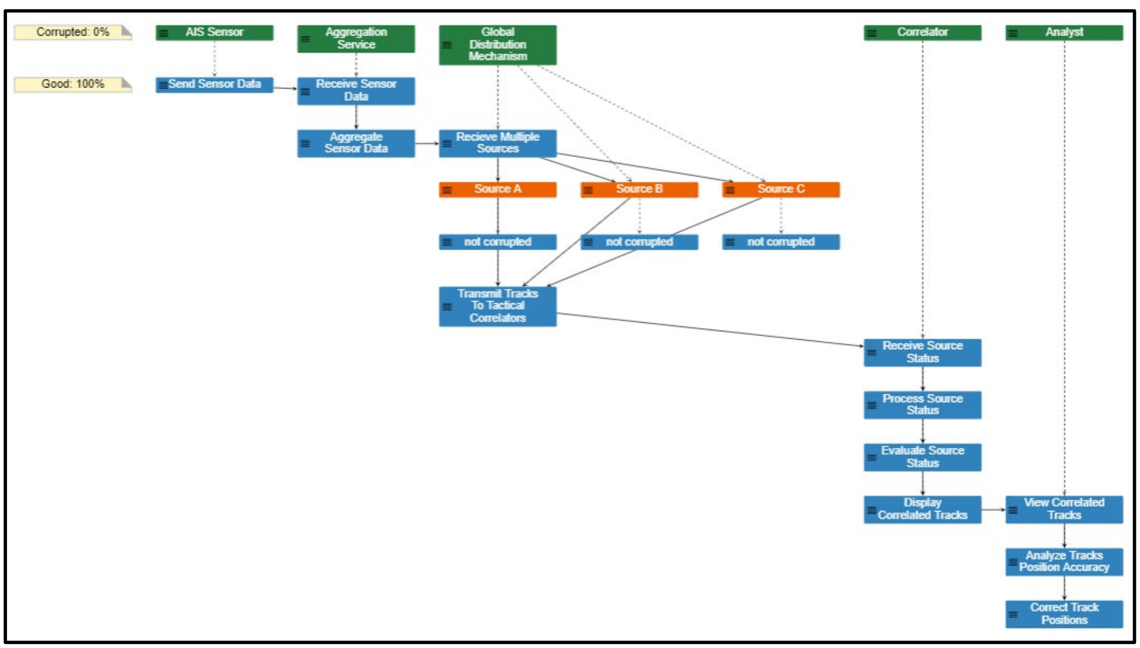

Figure 7. Model of Optimal Trace for Ungoverned Track Report Processing, Scope 1, Trace 1.

The key characteristic revealed in this model is the lack of any data source control methodology. This model shows the global distribution mechanism sending all data sources, unevaluated, to tactical track correlators for processing. This model demonstrates unintended persistence of corrupted data resulting from this global distribution mechanism.

2. Characterization of Events with Probability Attributes

Monterey Phoenix creates views of systems architectures as high-level description of possible system behaviors, while focusing on dependencies in behavior of subsystems (components) and interactions between them. For the purposes of this research, components are nodes representing functions of enterprise data flow of AIS reports into tactical correlators. Events in MP are abstract representations of activities (M. Auguston 2020, 5). When event traces are created, a probability of each event trace's occurrence out of the set of generated traces is automatically computed. Auguston calls this "Type 1 probability" (p. 58); it is shown as a p = *nnn* number above the thumbnail of each trace. MP users can also compute the probability of an event appearing *within* a particular event trace – a conditional probability; this is referred to by Auguston as "Type 2 probability" (p. 60). This research focuses on the Type 1 probabilities computed for the event traces in the MP schemas created for data correlation and fusion, since Type 1 is sufficient to gain the high-level insights about the unexpected systems design issues and unintended data flow behaviors associated with this multi-source, "big data" data architecture needed to answer the research question.

The Type 1 calculation process implemented by MP is broken down into the following steps. First, the probabilities assigned to events in each trace segment are multiplied to yield an unconstrained trace probability. Next, any trace segments that are rejected by constraints are removed and their associated probabilities are discarded. Finally, the probabilities of the remaining valid traces are normalized to add up to 1 by summing the probabilities of every valid trace and dividing each valid trace's probability by the sum of all the valid traces (Auguston 2020, 58).

All MP models in this thesis used the Type 1 probabilities based on the set of valid, finite, scope-complete behaviors, in order to track the probability of the event traces. The statistical data source evaluations presented in Figure 6 were used to assign realistic probabilities to events of the modeled data sources in MP. In order to represent real world track reporting conditions for sources A, B, and C, data sources A and B were given a data corruption percentage of 1% while data source C was set to 97% for modeling purposes. Table 2 summarizes the MP event trace results for ungoverned track report processing.

Table 2. Ungoverned Track Report Traces, Scope 1

Trace #	Optimal Data flow (True or False)	Type 1 Trace Probability	Data source corruption status
1	True	0.029403	No corrupted data
2	False	0.950697	Data Source A not corrupted Data Source B not corrupted Data Source C corrupted
3	False	0.000297	Data Source A not corrupted Data Source B corrupted Data Source C not corrupted
4	False	0.009603	Data Source A not corrupted Data Source B corrupted Data Source C corrupted
5	False	0.000297	Data Source A corrupted Data Source B not corrupted Data Source C not corrupted
6	False	0.009603	Data Source A corrupted Data Source B not corrupted Data Source C corrupted
7	False	0.000003	Data Source A corrupted Data Source B corrupted Data Source C not corrupted
8	False	0.000097	Data Source A corrupted Data Source B corrupted Data Source C corrupted

Trace 1 (Figure 7) represents receipt of no corrupted data (all data sources are good). The Type 1 probability of Trace 1 out of all 8 traces at scope 1 is only 2.94% (from Table 2). Since the sum of MP trace probabilities must equate to 1, subtracting the only uncorrupted trace from 1 provides the probability of all other traces that contain corrupted

data to occur. The probability for processing corrupted position reports from any of the data sources is therefore 97.06%:

$$\rho_{corrupted} = 1.0 - .0294 = .9706$$

The highest probability trace from all eight traces listed in Table 2 is shown in Figure 8, which presents a focused trace of the Global Distribution Mechanism showing the current method of AIS report distribution.

An instance of a corrupted report transmission is highlighted in red. The Type 1 probability of trace 2 is 95.07%. This trace results in data source C sending corrupt reports to tactical correlators.

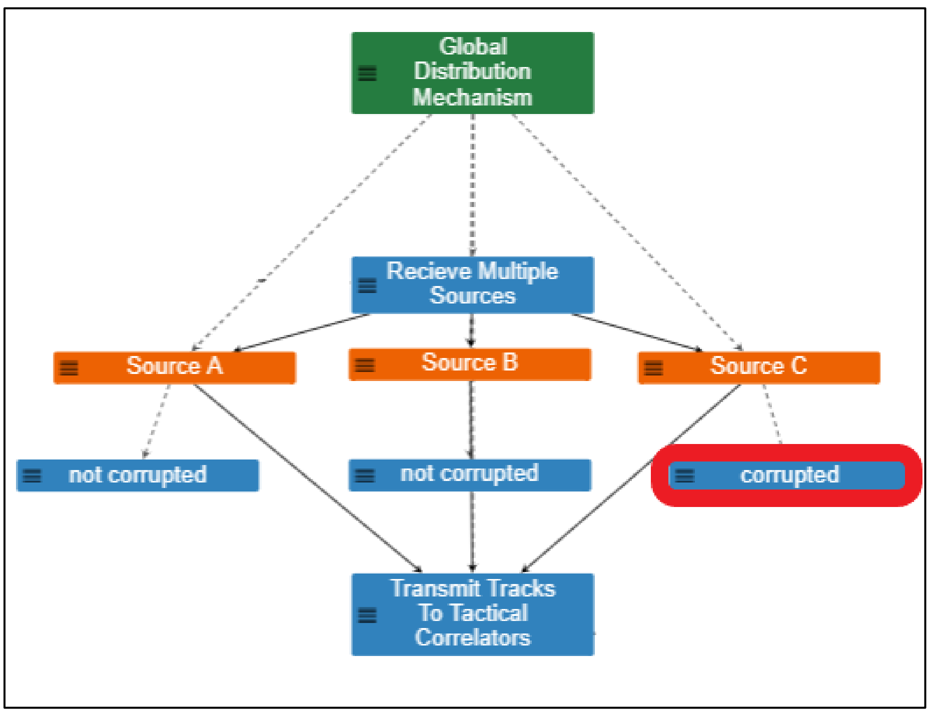

Figure 8. Most Probable Ungoverned Data Flow ($\rho =$ .9507), Scope 1, Trace 2

The probability of the trace showing all 3 data sources sending uncorrupted reports to tactical correlators (Figure 9) is only 2.94% according to Table 2, Trace 1.

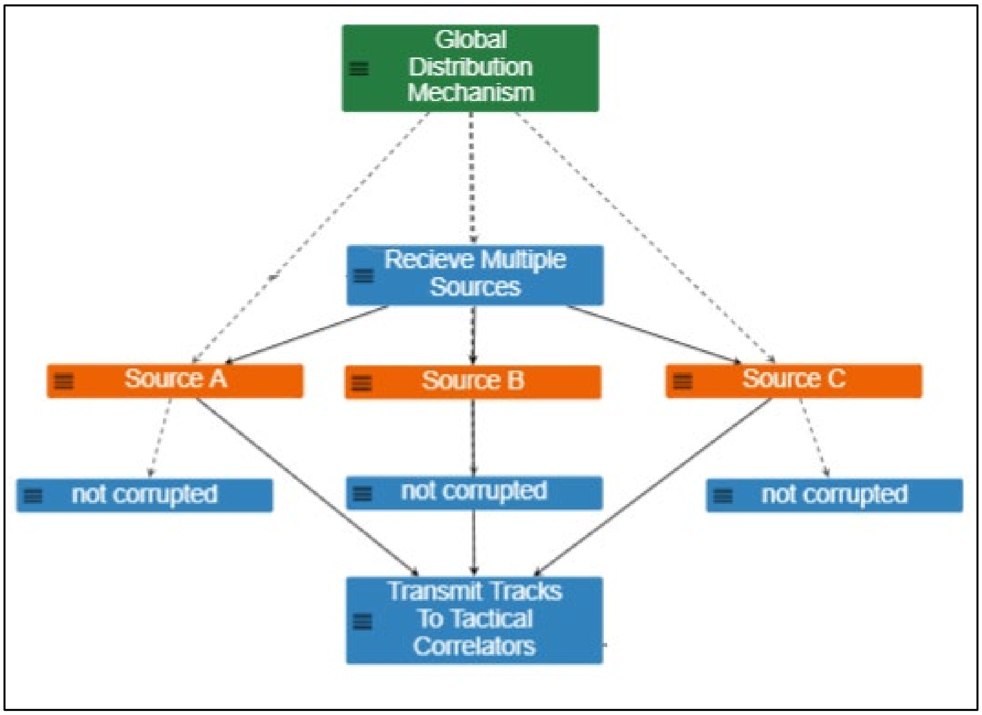

Figure 9. Event Trace Showing Uncorrupted Data Flow ($\rho = .0294$), Scope 1, Trace 1

The global report in Figure 10 was created in Monterey Phoenix. First, valid traces of interest that did not contain any AIS report corruption were accumulated and stored. Then the probabilities of those traces were summed and reported as uncorrupted (good) data sources. Finally, the summed probability of the good data sources was subtracted from 1, providing the probability in all valid traces for corrupted report data flow. The global report provides two probabilities. First, it provides the probability of an event trace transmitting good data. Finally, the report shows the probability of an event trace transmitting corrupted data.

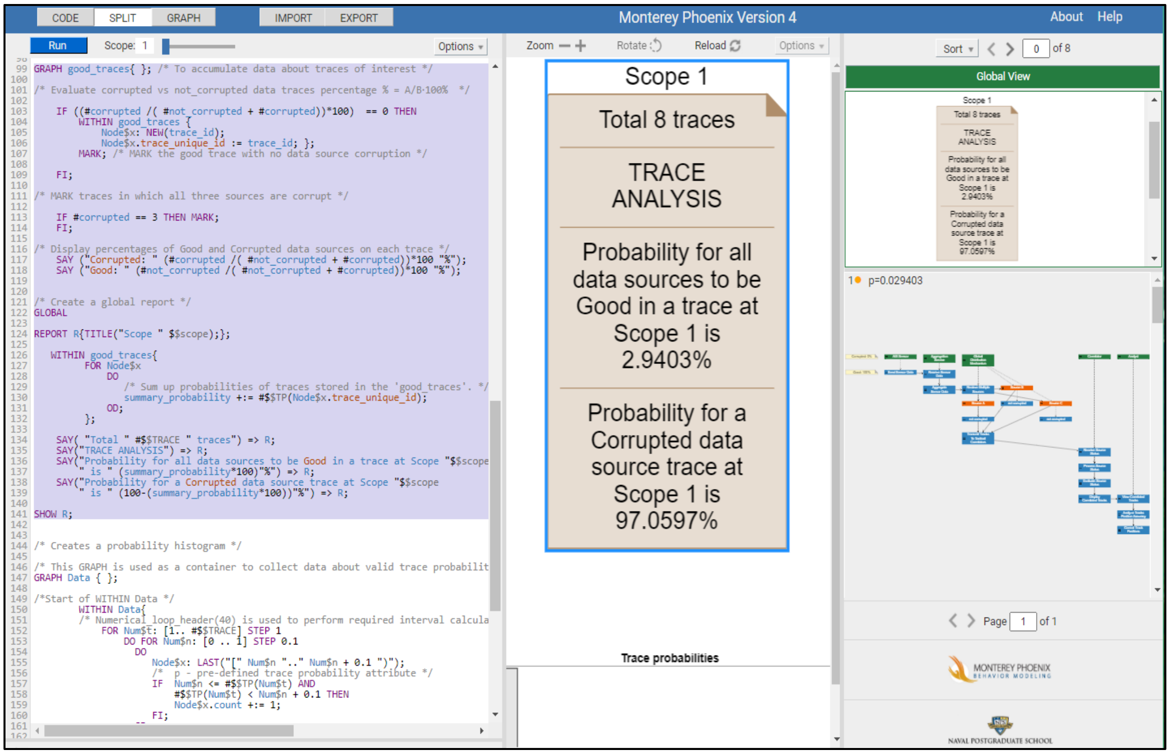

Figure 10. MP Firebird Tool with Ungoverned Behavior Global Report

Taken together, the event traces presented in this section demonstrate that current methods of battlespace correlation ignore substantial aberrant behaviors created by corrupted data sources, since all three track sources are transmitted for correlation regardless of the known accuracy properties of the tracks. The code for this ungoverned behavior model, as well as the associated global report is included in the Appendix.

D. MODELING WITH AN APPLIED DATA GOVERNANCE FRAMEWORK

Multi-source track correlation presents unique challenges in regards to the application of data governance. Processing position reports from uncontrolled, divergent sources presents data management problems, deliberately aggregating data from different silos of data that cannot be collectively controlled. In order to present a more accurate picture of the battlespace, multiple, diverse sources must be ingested, but if the sources are ingested without evaluation of corruption, validity, and accuracy, correlators are forced to ingest all data without accounting for futile ingestion of invalid data.

A data governance scoring rubric was created to generate a sound method of characterizing "rudimentary" data quality from each of the three AIS data sources. This scoring system, defined as a "checklist scoring rubric" was applied to data sources A, B, and C in Table 3. The rubric is designed to allow the majority of all AIS reports to flow into the correlator function while stopping or filtering flow for only unacceptable or emergent data sources. Six levels of rubric values are defined in the table. As detailed in the rubric legend, the "unacceptable" rubric score of 1 is applied to report data sources where 0-20% of all data is populated with valid data in critical report fields. The "emergent" data score of 2 is applied when valid data is present in 21-40% of all reports from that data source. The "minimally acceptable" data score of 3 is applied when valid data is present in 41-54% of all reports from a specific data source. An "acceptable" data score of 4 is applied to the data source when valid data is present in 55-79% of all reports from the data source. The "accurate" rubric score of 5 is applied to data sources where 80-94% of all reports contain valid data. Finally, the "highly accurate" score of 6 is applied to data sources where valid data is present in 95-100% of all reports from that data source.

Table 3. Discrimination Factor Rubric

Six Level Rubric (x represents actual % of valid source data): Unacceptable: Critical data fields present/valid in ($x < 21\%$) of source data. Emergent data: Critical data fields present/valid in ($21\% \leq x < 41\%$)of source data Minimally Acceptable: Critical data fields present/valid in ($41\% \leq x < 55\%$) of source data Acceptable: Critical data fields present/valid in ($55\% \leq x < 80\%$) of source data Accurate: Critical data fields present/valid in ($80\% \leq x < 95\%$) of source data Highly Accurate: Critical data fields present/valid in ($95\% \leq x < 100\%$) of source data		
Label	Rubric Value	Evidence based reasoning
Data Source A	6	Data source had no significant level of missing or corrupted data fields.
Data Source B	6	Data source had no significant level of missing or corrupted data fields.
Data Source C	1	High levels of missing fields and corrupt data were measured from this data source.

According to evidence-based reasoning from the existing statistical analysis, data source A received a rubric value of “6”, data source B received a value of “6”, while data source C received the lowest rubric score possible, “1”, as high levels of missing fields and corrupt data were measured from this data source.

The previous model that demonstrated data flow from sensor all the way to the human analyst was updated with a data governance framework that implements the discrimination factor rubric. The MP model was revised by adding a data governance framework. In the model, data flow is modeled starting at the sensor, flowing to the aggregation service, then routing to the distribution mechanism. All three USCG data sources are represented and transmitted to the data governance framework. The data sources are then evaluated with x representing the actual % of valid source data. If any of the AIS report sources received a statistical evaluation where $(x < 41\%)$, that source is filtered, and not sent to the correlator. Any report sources where $(x \geq 41\%)$ are scored a 3 or higher by the discrimination factor rubric. This evaluated source meets the criteria of “minimally acceptable data source”, and the data governance framework allows transmission of these reports to the correlator. These remaining, validated AIS reports are processed and evaluated in the correlator prior to being sent to the display for analyst viewing. The model was run at scope 1, producing 4 event traces for analysis. Figure 11 shows the applied data governance framework in the context of the total process.

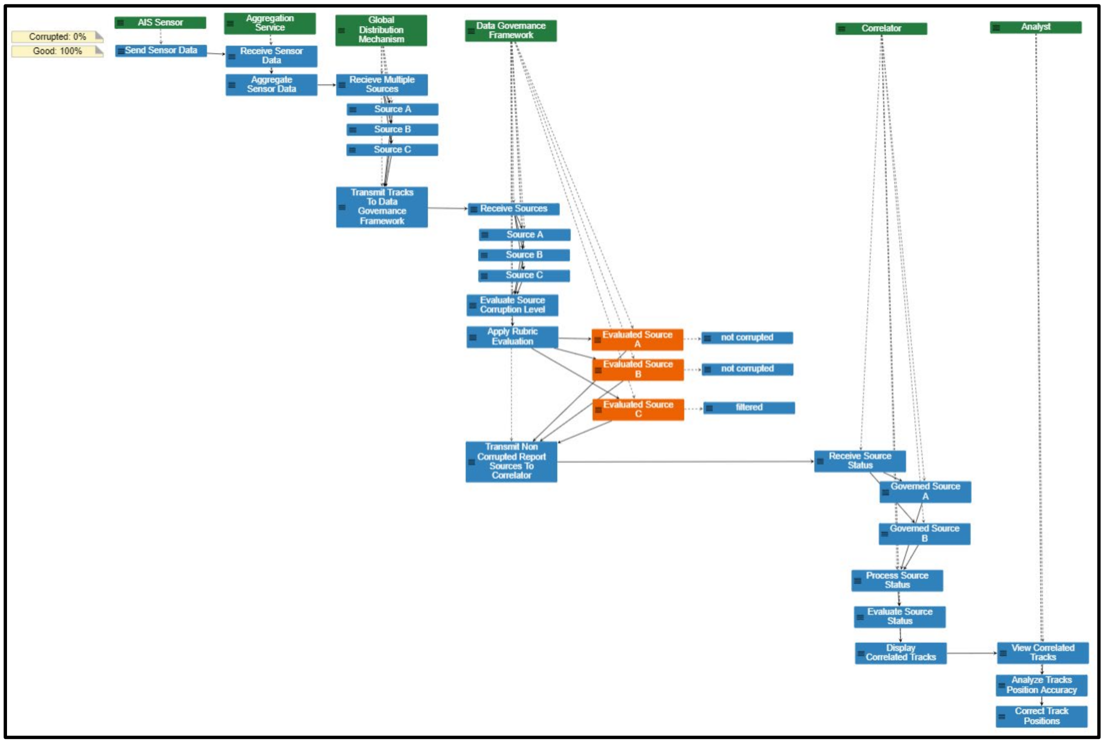

Figure 11. Applied Data Governance Framework, Scope 1, Trace 1

A focused section of the MP model is presented in Figure 12 to show the new root titled the "Data Governance Framework" in which the rubric evaluation takes place. The data governance framework diagram in Figure 12 is focused to demonstrate the applied data governance framework. Multiple sources are transmitted to the data governance framework function. The governance of the data sources is enforced via the "Apply Rubric Evaluation" function. Each individual source is received and evaluated independently, as this function evaluates reporting source corruption levels. Reporting sources A, B, and C are separately evaluated. While reporting Sources A, and B receive evaluations of "highly accurate", a rubric value of "6", the corruption levels of reporting source C result in an unacceptable rubric value of "1". The "Apply Rubric Evaluation" function allows report sources A and B to flow to the correlator. The allowed output flows display as "not corrupted". Data source C receives the "filter" status applied to its data and is prevented from being sent to the correlator. This results in an effective data governance framework that filters track reports that are unacceptable from being sent to the correlator for processing.

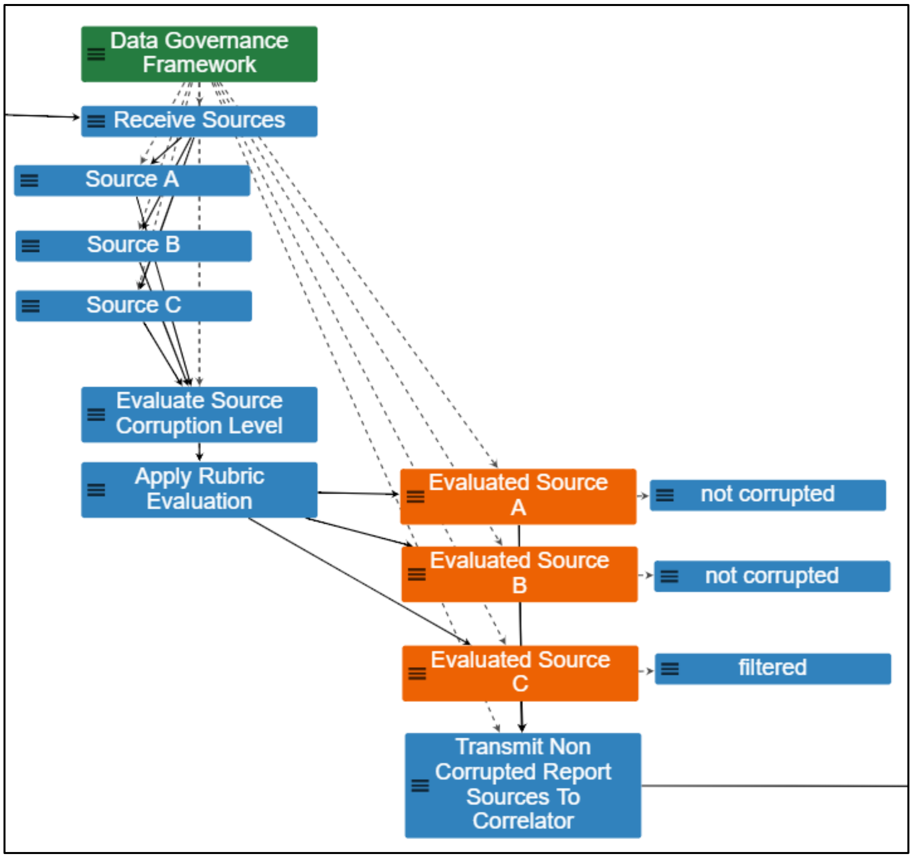

Figure 12. Focused Data Governance Framework, Scope 1, Trace 1

The enforced data governance framework results are displayed in Table 4. With data governance applied, data source C is stripped from all traces due to enforcement of the assigned rubric value of 1, as ($x < 21\%$) for all AIS reports provided from this source. In the updated model only AIS report sources for which ($x \geq 41\%$) are allowed to pass beyond the data governance framework to the correlator.

After implementing a simulated data governance framework in the MP model when using the same data sources, the highest probability is represented in trace 1, demonstrating a 98.01% probability of correlators receiving uncorrupted track report data for processing (Table 4). The probability for processing corrupted data is the sum of the remaining traces, only 1.99%.

Table 4. Applied Data Governance Framework Track Report Traces, Scope 1

Trace #	Optimal Data flow (True or False)	Type 1 Trace Probability	Data source corruption status
1	True	0.9801	No corrupted data
2	False	0.0099	Data Source A not corrupted Data Source B corrupted
3	False	0.0099	Data Source A corrupted Data Source B not corrupted
4	False	0.0001	Data Source A corrupted Data Source B corrupted

The global report in Figure 13 provides the probabilities of transmitting uncorrupted (good) and corrupted data in valid event traces when using the applied data governance framework model. The probabilities were calculated using the same method that was used for the ungoverned data model.

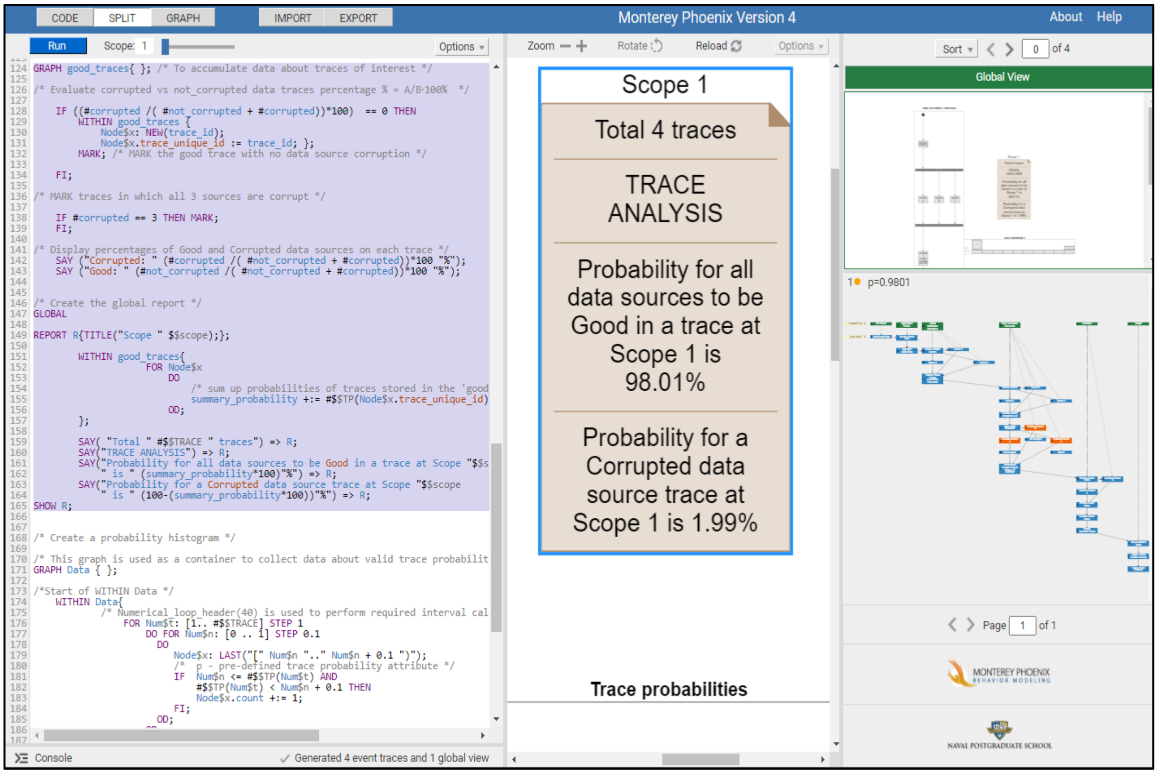

Figure 13. MP-Firebird Tool with Applied Data Governance Framework Global Report

A visualization of the data governance framework's event flow is provided in the activity diagram shown in Figure 14. It was created using a built-in command in the MP language for generating activity diagrams from grammar rules. The code is included in the Appendix, Figure 19, line 208. This activity flow shows the receipt of all three data sources, the evaluation of the data sources, application of the rubric, and the transmission of uncorrupted data sources to the correlator.

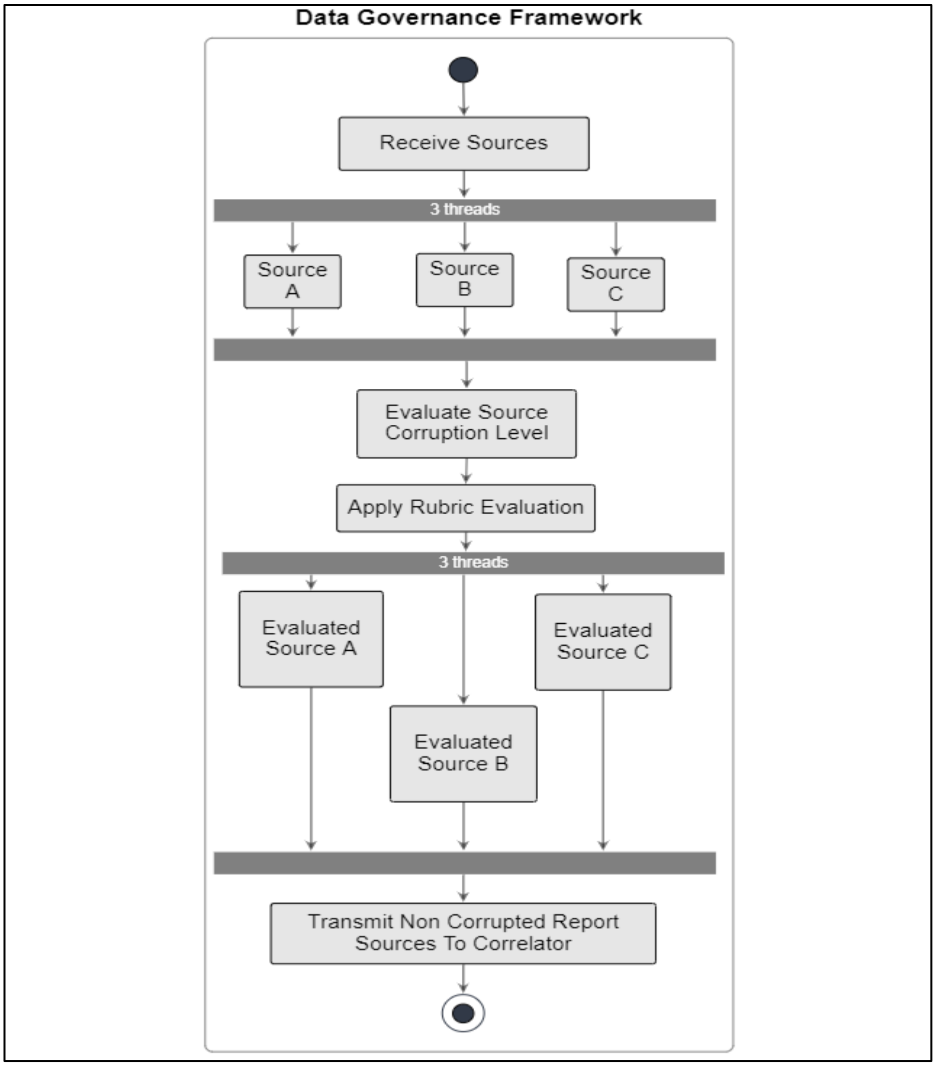

Figure 14. Activity Diagram of the Applied Data Governance Framework

The ungoverned behavior model and the applied data governance framework model were created and simulated using Monterey Phoenix to determine the prorated Type 1 trace probability of emergent, corrupted track report data flows into a correlator. The results are presented in a summary comparison of data source report corruption according to the combined global report graphic in Figure 15. The left portion of Figure 15 depicts two probabilities. The probability of a correlator processing verified AIS reports in valid event traces without any form of data governance is displayed, and the probability of corrupted data to be present in valid traces is presented in each report is displayed as well. The right side of the graphic displays the probability of a correlator processing verified AIS reports when a data governance framework is instantiated, while also displaying the probability of corrupted data to be present in valid traces in the lower area of the report.

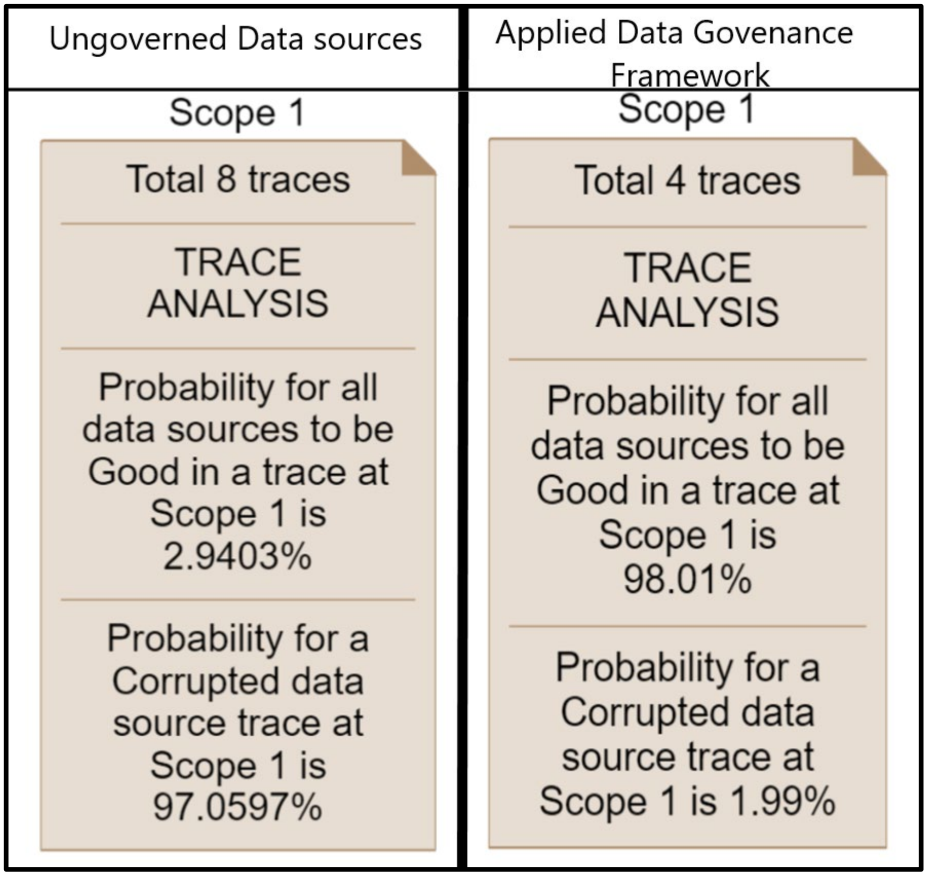

Figure 15. Global Reports for Ungoverned (left) and Governed (right) Data Sources

The Monterey Phoenix source code for all models and reports used in this research is included in the Appendix.

IV. CONCLUSIONS

This research contends that new systems engineering methodology and tools can help create more accurate depictions of the battlespace. This chapter presents conclusions based on the objectives identified in Chapter I, discusses research limitations and analysis, and provides future areas of research for systems engineers, analysts, and the greater of Maritime Domain Awareness (MDA) community.

A. RESEARCH CONCLUSIONS

Today, human analysts must expend significant effort and time to filter and discriminate unwanted data from correlator displays in order to create more accurate depictions of entities the battlespace. Any means of relieving unnecessary processing burden, reducing software processing latency and memory consumption will improve both correlator and human analyst performance.

The extreme volume of daily AIS track reports that are ingested today creates an immense correlator workload comparable to the extreme number of tasks associated with big data analysis. This workload includes processing corrupted position reports, in addition to evaluating all relevant position report data fields from good data sources, increasing the amount of valuable system computational resources for every report that is received. In light of this problem, this research asked the question, "How can behavior models of divergent data sources be used to design improvements in accuracy of battlespace object correlation and fusion?" A design and analysis research method was used to identify and explore algorithm-based multi-source fusion models, and to compare ungoverned and governed approaches to fusion. The research was conducted in three phases. The first phase modeled and evaluated current methodologies in use today. The second phase defined a new data governance-based approach to the problem, and the third phase integrated a hybrid methodology that implemented data governance, informing a determination about the extent results were improved.

The implementation of a formal modeling approach that demonstrates expected and unintended behaviors of complex system data flows provides unique insights not available via informal methods. This information provides a framework for implementing data governance, which has the potential to significantly relieve processing burden from correlation software while also reducing the burden of human analysts correcting correlated visualizations of ship positions.

Modeling event flows in Monterey Phoenix proved to have a value by itself, as it articulated both the current and future envisioned processes of data correlation and fusion in the context of the entire system (correlator) and the environment (data sources, human analysts). Modeling existing AIS report data flow using MP demonstrated divergent behaviors of ungoverned data sources, identifying high levels of AIS data source corruption. Creation of an AIS report data governance framework addressed this issue by applying discrimination factors evaluating the quality of unique data sources prior to correlation. This modeling provided the unique ability to create and analyze data governance framework trace results demonstrating the value of formal behavioral modeling of data sources that contain statistical variance. Tactical information flow modeling enabled the creation of a data governance framework that implemented applied discrimination factors, or "rubrics" for data sources prior to processing. The implementation of a data governance framework that discriminates and filters corrupted data from position report sources prior to sending those reports to an AIS correlator decreased the probability of tactical correlators processing corrupted AIS position reports.

The analyses in this research established that sending large quantities of corrupt position reports into AIS correlator software represents "futile" processing cycles in tactical correlation software. The formal models developed in Monterey Phoenix helped articulate the concept of a governance framework, and quantitatively demonstrated the impact to the accuracy of the fused picture of self-reporting AIS vessels in the battlespace. In this research, the impact was significant. The current, unregulated data architecture in use today provides only a 2.94% probability (based on the data set used) that all AIS reports sent to be correlated were completely usable, or uncorrupted. The implementation of a data governance framework on the same data set significantly increased the probability of completely usable, or uncorrupted AIS reports being sent for processing, increasing that probability from only 2.94%, to 98.01%.

This research validates the essential value of data governance and demonstrates the application of a data governance framework as a simple, viable, method of discriminating data source quality. Tactical correlation software runs more efficiently, decreasing resource consumption while providing a more accurate depiction of the battlespace, increasing the consistency and accuracy of data sources. Future correlation software and human analysts will be more efficient when receiving uncorrupted data. Human analysts will be able to invest more time and mental focus into examining entities and interactions vs. expending energy and focus using historical context and other methods to purge incorrect position reports from their three-dimensional understanding of the battlespace.

B. LIMITATIONS

Time constraints limited the scope of data analysis for this research. These constraints prevented in depth, statistical modeling of individual report field deviations from different data sources. For this reason, only Type 1 Monterey Phoenix model traces were used in this research as report data source corruption probabilities were not dependent on other data source events. Additionally, only general principles of the data governance framework were used to illustrate Type 1 probability. Data governance frameworks are capable of in-depth evaluation and screening of data sources but require detailed characterization of the data. Had time not been a constraint, this research would have implemented a much more robust MP modeling schema instantiating both Type 1 and Type 2 probabilities of event behaviors.

This research discovered a lack of existing peer reviewed academic literature focused on application of data governance in the field (MDA). While there are numerous reporting data standards, such as Tactical Digital Information Link, (TADIL) standard messages and Over The Horizon-Gold, (OTH-Gold) messaging formats, the community has not implemented data governance of this standardized data. To overcome lack of existing academic material, this research was forced to integrate concepts from big data analysis, and data governance definitions from NIST.

C. FUTURE WORK

Systems engineering research should conduct more advanced formal modeling with in depth assessments of data source field quality and accuracy. Using Type 2 probabilities could provide additional insight into future data governance frameworks by adding qualitative likelihoods of event occurrence that could be used to effectively prune data sources of bad data; these probabilities could also build impact factors depending on the needs of the analyst fusing the data. Impact factors could then be assigned to provide the likelihood of events occurring, effectively predicting risks to a system. Global risk reports could provide unique insight into active data architectures highlighting areas of system data flow that need immediate attention.

Future research could create a fielded software application leveraging Monterey Phoenix's capability to model a system and provide both Type 1 and Type 2 event probabilities while implementing an applied data governance framework for position reports. This enhanced software package could create a system of source evaluation rubrics and measures. Eventually, the software could include machine learning algorithms that detect emerging data anomalies in report sources, alerting analysts to failed data sources, or spoofed sensor data. Using this framework, the software could integrate historical databases to discern historical patterns of data deviation and corruption in data sources. Once data source deviations get detected, analysts will be able to analyze for any trends in the data, examining the data for aberrant patterns, helping to determine causes of the symptoms.

This research is applicable to many other data sources and types of data in addition to AIS position reports. There are many standardized communication methods in the Navy and commercial industry today. Any military, or civilian standardized communication reporting system could benefit from this research. Example message types would include but not be limited to Tactical Digital Information Link, (TADIL) standard messages, and Over The Horizon-Gold, (OTH-Gold) standard messages. Formally modeling enterprise data flows from the sensor to the analyst, with the addition of an applied data governance framework would optimize any standardized reporting distribution method used in the military and industry today. Performing these functions converts data source fidelity into valued information, which human analysts turn into knowledge. Additionally, evaluating data sources provides insightful, context-based information. This context helps define whether the data reported is reasonable. The MDA community could then ask the following knowledge management questions:

- What causes data source fidelity issues?
- Why is there corruption or degradation of a specific data source?
- Can analysts learn anything from data source trends?

Using data governance frameworks, MDA analysts could automatically evaluate future sensor networks and data distribution systems. If increasing data corruption levels and deviations get detected, source system maintainers could conduct preventive measures and reduce data corruption or correct data deviation trends such as time delays from inaccurate clocking or reducing distribution backlog issues. In addition, advanced applied data governance frameworks could provide constant health and status of national, regional, and local data sources. This capability could provide an automated means of data source status reporting, which is critical in the burgeoning field of predictive condition-based maintenance.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX. MONTEREY PHOENIX CODE

The ungoverned track report processing model and the applied data governance framework model were created and simulated using Monterey Phoenix. The complete modeling code is included for reference and divided into three different figures for each model.

```
/* Ungoverned Correlation Model August 2021
    created by F Watson 8/20/2021

    Purpose:
    To model AIS reporting data flows showing existing data path behaviors

    Description:
    Models the data flow from sensors all the way to the human analyst at a tactical location.

     Details
     AIS_Sensor: represents an example AIS GPS SENSOR
     Aggregation_Service: Represents a local AIS Aggregator like a "Radar" system or local AIS Broadcast
     Global_Distribution_Mechanism: Receives data from multiple aggregation services, coalates the data,
     and rebroadcasts the data globally.

     Source A: Represents raw AIS tracks that have not been modified in any way prior to dissemination
     Source B: Represents raw AIS tracks grouped into geographic areas for limited broadcast
     Source C: Represents Custom filtering and combining of AIS tracks based on unique fields
     Correlator functions:      Receive_Source_Status
                        Process_Source_Status
                        Evaluate_Source_Status
                    Display_Correlated_Tracks
     Analyst functions:         View_Correlated_Tracks
                   Analyze_Tracks_Position_Accuracy
                   Correct_Track_Positions

    The events are representative of actual AIS data report flow to operational analysts in the Fleet today.

    References:

    Example 49 Histogram showing number of traces with probabilities within certain intervals, from Auguston, M. "Monterey
    Phoenix System and Software Architecture and Workflow Modeling Language Manual"
    (Version 4). 2020. Available online: https://wiki.nps.edu/display/MP/Documentation

    Example 40. A global report is assembled from the set of all available event traces, from Auguston, M. "Monterey Phoenix
    System and Software Architecture and Workflow Modeling Language Manual"
    (Version 4). 2020. Available online: https://wiki.nps.edu/display/MP/Documentation

    Search terms: behavior, data correlation and fusion; probability, Type 1; data governance

    Instructions:
    Run for scope 1: 8 traces in < 1 sec



*/

SCHEMA Ungoverned_Correlation_Model


ROOT    AIS_Sensor: Send_Sensor_Data ;

ROOT    Aggregation_Service:    Receive_Sensor_Data
                                Aggregate_Sensor_Data ;

ROOT    Global_Distribution_Mechanism:  Recieve_Multiple_Sources
                                        { Source_A, Source_B, Source_C }
                                        Transmit_Tracks_To_Tactical_Correlators ;
```

Figure 16. Ungoverned Behavior Model Page-1

```
    Source_A:    (not_corrupted |<<0.01>> corrupted) ;
    Source_B:    (not_corrupted |<<0.01>> corrupted) ;
    Source_C:    (not_corrupted |<<0.97>> corrupted) ;

ROOT    Correlator:     Receive_Source_Status
                        Process_Source_Status
                        Evaluate_Source_Status
                        Display_Correlated_Tracks ;


ROOT    Analyst:    View_Correlated_Tracks
                    Analyze_Tracks_Position_Accuracy
                    Correct_Track_Positions ;


/* Model the data flow to from sensor to the Analyst/Operator */
    COORDINATE  $x: Send_Sensor_Data            FROM AIS_Sensor,
                $y: Receive_Sensor_Data         FROM Aggregation_Service
        DO  ADD $x PRECEDES $y; OD;

    COORDINATE  $x: Aggregate_Sensor_Data       FROM Aggregation_Service,
                $y: Recieve_Multiple_Sources    FROM Global_Distribution_Mechanism
        DO  ADD $x PRECEDES $y; OD;

    COORDINATE  $x: Transmit_Tracks_To_Tactical_Correlators FROM Global_Distribution_Mechanism,
                $y: Receive_Source_Status                   FROM Correlator
        DO  ADD $x PRECEDES $y; OD;

    COORDINATE  $x: Display_Correlated_Tracks   FROM Correlator,
                $y: View_Correlated_Tracks      FROM Analyst
        DO  ADD $x PRECEDES $y; OD;

/* Establish a trace count/start the count */
ATTRIBUTES {number  trace_unique_id,
                    summary_probability,
                    count; };

GRAPH good_traces{ }; /* To accumulate data about traces of interest */

/* Evaluate corrupted vs not_corrupted data traces percentage % = A/B·100%  */

    IF ((#corrupted /( #not_corrupted + #corrupted))*100)  == 0 THEN
        WITHIN good_traces {
            Node$x: NEW(trace_id);
            Node$x.trace_unique_id := trace_id; };
        MARK; /* MARK the good trace with no data source corruption */

    FI;

/* MARK traces in which all three sources are corrupt */

    IF #corrupted == 3 THEN MARK;
    FI;

/* Display percentages of Good and Corrupted data sources on each trace */
    SAY ("Corrupted: " (#corrupted /( #not_corrupted + #corrupted))*100 "%");
    SAY ("Good: " (#not_corrupted /( #not_corrupted + #corrupted))*100 "%");


/* Create a global report */
```

Figure 17.   Ungoverned Behavior Model Page-2

```
GLOBAL

REPORT R{TITLE("Scope " $$scope);};

   WITHIN good_traces{
         FOR Node$x
             DO
                /* Sum up probabilities of traces stored in the 'good_traces'. */
                summary_probability +:= #$$TP(Node$x.trace_unique_id);
             OD;
         };

    SAY( "Total " #$$TRACE " traces") => R;
    SAY("TRACE ANALYSIS") => R;
    SAY("Probability for all data sources to be Good in a trace at Scope "$$scope
        " is " (summary_probability*100)"%") => R;
    SAY("Probability for a Corrupted data source trace at Scope "$$scope
        " is " (100-(summary_probability*100))"%") => R;

SHOW R;


/* Creates a probability histogram */

/* This GRAPH is used as a container to collect data about valid trace probabilities */
GRAPH Data { };

/*Start of WITHIN Data */
        WITHIN Data{
        /* Numerical_loop_header(40) is used to perform required interval calculations. */
            FOR Num$t: [1.. #$$TRACE] STEP 1
                DO FOR Num$n: [0 .. 1] STEP 0.1
                  DO
                     Node$x: LAST("[" Num$n ".." Num$n + 0.1 ")");
                     /*  p - pre-defined trace probability attribute */
                     IF  Num$n <= #$$TP(Num$t) AND
                         #$$TP(Num$t) < Num$n + 0.1 THEN
                         Node$x.count +:= 1;
                     FI;
                  OD;
                OD;
        }; /* end of WITHIN Data */

/* Formats the probability chart */
TABLE probability_histogram {
                        TABS string     probability_interval,
                             number     trace_count;
                            };

                        BAR CHART probability_chart { TITLE("Trace probabilities");
                                                      FROM     probability_histogram;
                                                      X_AXIS probability_interval;
                                                    };
                        WITHIN Data{
                                    FOR Node$n
                                    DO probability_histogram     <|
                                             probability_interval:     SAY(Node$n),
                                             trace_count:              Node$n.count;
                                    OD;
                                    };

SHOW probability_chart SORT;
/* The contents of the chart is sorted by X_AXIS string values, since the order of adding rows to the
    TABLE may be arbitrary */
```

Figure 18. Ungoverned Behavior Model Page-3

```
/* Applied Data Governance Framework Model
created by F Watson 8/20/2021

    Purpose:
    To model AIS reporting data flows showing updated data path behaviors with the application of a data governance
    framework

    Description:
    Models the data flow from sensors all the way to the human analyst at a tactical location incorporating a data
    governance framework.

     Details
     AIS_Sensor: represents an example AIS GPS SENSOR
     Aggregation_Service: Represents a local AIS Aggregator like a "Radar" system or local AIS Broadcast
     Global_Distribution_Mechanism: Receives data from multiple aggregation services, coalates the data, and rebroadcasts
     the data globally.

     Source A: Represents raw AIS tracks that have not been modified in any way prior to dissemination
     Source B: Represents raw AIS tracks grouped into geographic areas for limited broadcast
     Source C: Represents Custom filtering and combining of AIS tracks based on unique fields

     Data Governance Framework applies evaluation rubric to data sources and only allows sources that are uncorrupted to be
     transmitted to the correlator.
     The Data_Governance_Framework receivees sources, evaluates source corruption levels, applies a rubric evaluation, and
     transmits approved sources to the correlator function


     Correlator functions:      Receive_Source_Status
                        Process_Source_Status
                        Evaluate_Source_Status
                    Display_Correlated_Tracks
     Analyst functions:         View_Correlated_Tracks
                   Analyze_Tracks_Position_Accuracy
                   Correct_Track_Positions

    The events are representative of actual AIS data report flow to operational analysts, with the insertion of a data
    governance framework.

    References:

    Example 49 Histogram showing number of traces with probabilities within certain intervals, from Auguston, M. "Monterey
    Phoenix System and Software Architecture and Workflow Modeling Language Manual"
    (Version 4). 2020. Available online: https://wiki.nps.edu/display/MP/Documentation

    Example 40. A global report is assembled from the set of all available event traces, from Auguston, M. "Monterey Phoenix
    System and Software Architecture and Workflow Modeling Language Manual"
    (Version 4). 2020. Available online: https://wiki.nps.edu/display/MP/Documentation

    Search terms: behavior, data correlation and fusion; probability, Type 1; data governance

    Instructions:
    Run for scope 1: 4 traces in < 1 sec


*/
SCHEMA Applied_Data_Governance_Framework_Model



ROOT    AIS_Sensor:     Send_Sensor_Data ;

ROOT    Aggregation_Service:    Receive_Sensor_Data
                                Aggregate_Sensor_Data ;

ROOT    Global_Distribution_Mechanism:  Recieve_Multiple_Sources
                                        { Source_A, Source_B, Source_C }
                                        Transmit_Tracks_To_Data_Governance_Framework ;

ROOT    Data_Governance_Framework:  Receive_Sources
                                    { Source A, Source B, Source C }
```

Figure 19. Applied Data Governance Framework-1

```
                                      Evaluate_Source_Corruption_Level
                                      Apply_Rubric_Evaluation
                                      {   Evaluated_Source_A,
                                          Evaluated_Source_B,
                                          Evaluated_Source_C }

                                      Transmit_Non_Corrupted_Report_Sources_To_Correlator ;


                    Evaluated_Source_A:      (not_corrupted |<<0.01>> corrupted) ;
                    Evaluated_Source_B:      (not_corrupted |<<0.01>> corrupted) ;

                    /*Data Source C receives a "filtered" status from the data governance framework*/
                    Evaluated_Source_C:      filtered ;


ROOT    Correlator: Receive_Source_Status
                    { Governed_Source_A, Governed_Source_B }
                    Process_Source_Status
                    Evaluate_Source_Status
                    Display_Correlated_Tracks ;

ROOT    Analyst:    View_Correlated_Tracks
                    Analyze_Tracks_Position_Accuracy
                    Correct_Track_Positions ;

/* Model the data flow to from sensor to the Analyst/Operator */
COORDINATE  $x: Send_Sensor_Data                FROM AIS_Sensor,
            $y: Receive_Sensor_Data             FROM Aggregation_Service
    DO  ADD $x PRECEDES $y; OD;

COORDINATE  $x: Aggregate_Sensor_Data           FROM Aggregation_Service,
            $y: Recieve_Multiple_Sources        FROM Global_Distribution_Mechanism
    DO  ADD $x PRECEDES $y; OD;

COORDINATE  $x: Transmit_Tracks_To_Data_Governance_Framework    FROM Global_Distribution_Mechanism,
                $y: Receive_Sources                             FROM Data_Governance_Framework
    DO  ADD $x PRECEDES $y; OD;

COORDINATE  $x: Transmit_Non_Corrupted_Report_Sources_To_Correlator FROM Data_Governance_Framework,
                $y: Receive_Source_Status                           FROM Correlator
    DO  ADD $x PRECEDES $y; OD;

COORDINATE  $x: Display_Correlated_Tracks               FROM Correlator,
                $y: View_Correlated_Tracks              FROM Analyst
    DO  ADD $x PRECEDES $y; OD;

/* Establish a trace count/start the count */
ATTRIBUTES {number  trace_unique_id,
                    summary_probability,
                    count; };

GRAPH good_traces{ }; /* To accumulate data about traces of interest */

/* Evaluate corrupted vs not_corrupted data traces percentage % = A/B·100%  */

    IF ((#corrupted /( #not_corrupted + #corrupted))*100)  == 0 THEN
        WITHIN good_traces {
            Node$x: NEW(trace_id);
            Node$x.trace_unique_id := trace_id; };
        MARK; /* MARK the good trace with no data source corruption */

    FI;

/* MARK traces in which all 3 sources are corrupt */

    IF #corrupted == 3 THEN MARK;
    FI;
```

Figure 20. Applied Data Governance Framework-2

```
/* Display percentages of Good and Corrupted data sources on each trace */
    SAY ("Corrupted: " (#corrupted /( #not_corrupted + #corrupted))*100 "%");
    SAY ("Good: " (#not_corrupted /( #not_corrupted + #corrupted))*100 "%");


/* Create the global report */
GLOBAL

REPORT R{TITLE("Scope " $$scope);};

        WITHIN good_traces{
                    FOR Node$x
                        DO
                            /* sum up probabilities of traces stored in the 'good_traces'. */
                            summary_probability +:= #$$TP(Node$x.trace_unique_id);
                        OD;
        };

        SAY( "Total " #$$TRACE " traces") => R;
        SAY("TRACE ANALYSIS") => R;
        SAY("Probability for all data sources to be Good in a trace at Scope "$$scope
            " is " (summary_probability*100)"%") => R;
        SAY("Probability for a Corrupted data source trace at Scope "$$scope
            " is " (100-(summary_probability*100))"%") => R;
SHOW R;

/* Create a probability histogram */

/* This graph is used as a container to collect data about valid trace probabilities */
GRAPH Data { };

/*Start of WITHIN Data */
    WITHIN Data{
            /* Numerical_loop_header(40) is used to perform required interval calculations. */
                FOR Num$t: [1.. #$$TRACE] STEP 1
                    DO FOR Num$n: [0 .. 1] STEP 0.1
                      DO
                         Node$x: LAST("[" Num$n ".." Num$n + 0.1 ")");
                         /*  p - pre-defined trace probability attribute */
                         IF  Num$n <= #$$TP(Num$t) AND
                             #$$TP(Num$t) < Num$n + 0.1 THEN
                             Node$x.count +:= 1;
                         FI;
                      OD;
                    OD;
                }; /* end of WITHIN Data */

/* Format Probabilty Chart */
TABLE probability_histogram {
                             TABS string     probability_interval,
                                  number     trace_count;
                           };

BAR CHART probability_chart { TITLE("Trace probabilities");
                             FROM     probability_histogram;
                             X_AXIS probability_interval;
                           };
        WITHIN Data{
                    FOR Node$n
                        DO probability_histogram     <|
                                probability_interval:     SAY(Node$n),
                                trace_count:              Node$n.count;
                        OD;
                    };

    SHOW probability_chart SORT;
/* The contents of the chart are sorted by X_AXIS string values*/
SHOW ACTIVITY DIAGRAM Data_Governance_Framework;
```

Figure 21. Applied Data Governance Framework-3

LIST OF REFERENCES

Auguston, Mikhail and Kristin Giammarco. 2019. "Monterey Phoenix Tutorial." May. Accessed April 15th, 2021. https://wiki.nps.edu/display/MP/Tutorials?preview=%2F1036681365%2F1163467194%2FMP+Tutorial+-+Abridged+PDFs.zip.

Auguston, Mikhail. 2020. Monterey Phoenix System and Software Behavior Modeling Language 4.0. Monterey, CA, March 03.

Collins Dictionary. 2021. Definition of 'battlespace'. Accessed May 23, 2021. https://www.collinsdictionary.com/us/dictionary/english/battlespace.

Dietrich, Nathan S. 2001. Performance metrics for correlation and tracking algorithms. June. Accessed May 23, 2021. https://calhoun.nps.edu/handle/10945/2473.

Fernandez Arguedas, Virginia, Giuliana Pallotta, and Michele Vespe. 2018. "Maritime Traffic Networks: From Historical Positioning Data to Unsupervised Maritime Traffic Monitoring." IEEE Transactions on Intelligent Transportation Systems, vol. 19, March 722–732.

Giammarco, Kristin. 2021. "Model of Stack Behavior," MP Model Collection, . Sept 3. Accessed Sept 3, 2021. https://gitlab.nps.edu/monterey-phoenix/mp-model-collection/preloaded-examples/-/tree/master/models.

Henk, Blom and Edwin Bloem. 2006. "Exact Bayesian filter and joint IMM coupled PDA tracking of maneuvering targets from possibly missing and false measurements,." *Automatica*, Issue 1 127-135.

Hodson, Eric. 1975. "Some Exponential Alternatives to the Normal Model in Time Series Analysis." Thesis. Monterey, CA: Naval Postgraduate School, June. Accessed January 22nd, 2021.

Langley, Richard B. 1999. "Dilution of Precision." GPS World 54.

Liu, Weiming, Chen Zhang, Bin Yu, and Yitong Li. 2019. "A General Multi-Source Data Fusion Framework." ACM digital library. February. Accessed May 23, 2021. https://doi.org/10.1145/3318299.3318394.

National Institute of Standards. 2019. "NIST.SP.1500-10r1." doi.org. Oct. Accessed April 15th, 2021. https://doi.org/10.6028/NIST.SP.1500-10r1.

Perobelli, Nicola. 2016. A-day-in-numbers. June 28th. Accessed May 25, 2021. https://www.marinetraffic.com/blog/a-day-in-numbers/.

Quartuccio, John J. and Kristin M. Giammarco. 2018. *Engineering Emergence*. New York: CRC Press.

Rothenhaus, Kurt Joseph. 2008. “Data strategies to support automated.” NPS Institutional Archive. June. Accessed May 14th, 2021. https://calhoun.nps.edu/handle/10945/10348.

United States Coast Guard. 2021. Navigation Center. Feb 19th. Accessed April 15th, 2021. https://www.navcen.uscg.gov/?pageName=AISMessagesA#RAIM.

Wilner D. 1976. “Kalman Filter Algorithms for a multi-sensor system.” doi.org. Accessed May 14th, 2021. https://doi.org/10.1109/CDC.1976.267794.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California